BXRL Behavior-Explainable Reinforcement Learning

Center for Human-Compatible Artificial Intelligence

Ram Rachum[1,2] Yotam Amitai[3] Yonatan Nakar[4] Reuth Mirsky[2,†] Cameron Allen[1,†]

[1]UC Berkeley [2]Tufts University [3]Independent Researcher [4]Google Research [†]Equal senior contribution

GOLD Lab

RL agents mess up sometimes, and operators want to investigate why. XRL methods explain a single action at a specific point in time; we propose developing XRL methods that explain recurring patterns of actions specified by the operator.

We adapt a definition for **behavior** from the evolutionary-computation literature, originating with Lehman and Stanley [1]:

> **Definition 1 (Behavior measure).** Given an MDP, a behavior measure is any function m from a policy π to a scalar:
>
> $$m : \Pi \to \mathbb{R}$$
>
> A larger value of $m(\pi)$ indicates that π exhibits the represented behavior more strongly.

Examples of behavior measures in different domains:

Domain	Behavior	Metric
IPD	Reciprocity	$m(\pi) := \mathbb{E}_{o\sim D}[\pi_\theta(a^{\text{opp}}_{t-1} \mid o)]$
Driving	Tailgating	$m(\pi) := \mathbb{E}_{o\sim D_{\text{close}}}[1 - \pi_\theta(\text{slower} \mid o)]$
LLMs	Sycophancy	$m(\pi) := \mathbb{E}_{x\sim D}[\pi_\theta(\text{“And honestly— That's growth ✨”} \mid x)]$
Robotics	Jerkiness	$m(\pi) := \mathbb{E}_{(o,o')\sim D_{\text{consec}}}\left[\lVert\mu_\theta(o') - \mu_\theta(o)\rVert^2\right]$
Trading	Risk appetite	$m(\pi) := \mathbb{E}_{o\sim D}[\lvert\mu_\theta(o)\rvert]$
Recommenders	Popularity bias	$m(\pi) := \mathbb{E}_{o\sim D}\left[\sum_{i\in \text{ top-}k} \pi_\theta(i \mid o)\right]$

Table 1. Examples of expressing behaviors in different domains as a measure $m(\pi)$. D is a fixed observation distribution, independent of θ, ensuring the full computation from θ to $m(\pi_\theta)$ is end-to-end differentiable. Subscripted variants (D_{close}, D_{consec}) restrict D to scoped situations. Practitioners may draw observations from rollouts of a reference policy, a replay buffer, or construct them synthetically.

No XRL method currently allows users to define a behavior and receive an explanation for it; no XRL survey defines an expression for behavior. We provide a cross-taxonomy of explanation targets across XRL:

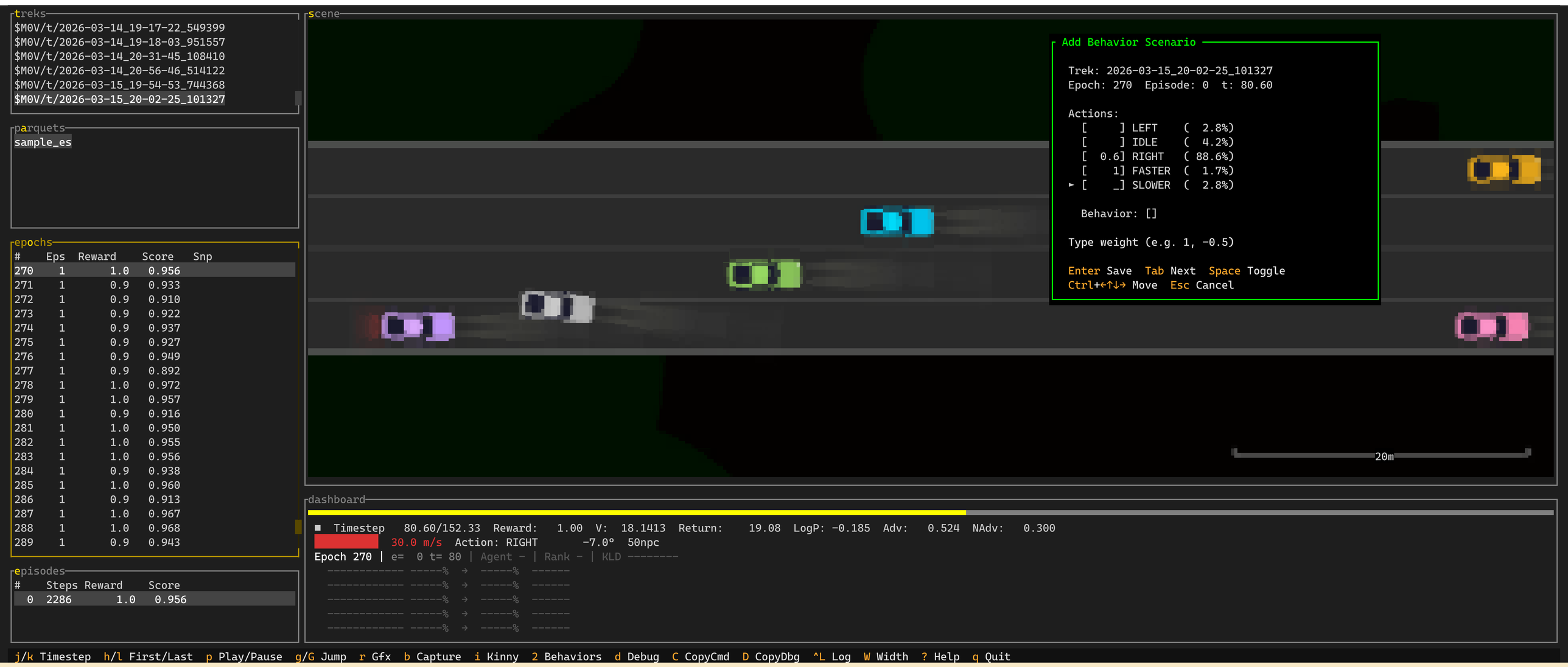
Explanation target	Example question	Vouros [2]	Milani et al. [3]	Saulières et al. [4]	# papers, exemplars
Action	“In episode 137, timestep 82, why did the agent choose action a?”	Outcome	Feature Importance: Directly Generate Explanations	Action	105 [5, 6]
Trajectory	“In episode 137, which timesteps were critical for the outcome?”	—	—	Sequence	11 [7, 8]
Policy	“What is the agent’s general decision-making process?”	Policy	Feature Importance: Interpretable Policies, Policy-Level	Policy	192 [9, 10]
Objective	“What is the agent optimizing? What are its goals?”	Objectives	—	—	3 [11, 12]
Learning	“Why did the policy gradually evolve to be the way that it is?”	—	Learning Process and MDP	—	17 [13, 14]

Table 2. A cross-taxonomy of XRL methods according to their explanation targets. The teal cells indicate the corresponding categories in three surveys.

We show how three existing explainability methods [15, 16, 17] may be adapted for BXRL. We provide an implementation of the `highway-v0` driving environment [18] in JAX with a TUI that allows defining and measuring behaviors, seen below. We define an example behavior that’s correlated with collisions, and call for the XRL community to develop methods that target this behavior.

Paper: r.rachum.com/bxrl-pdf

Code: github.com/HumanCompatibleAI/HighJax



[1] J. Lehman and K. O. Stanley. Abandoning Objectives: Evolution Through the Search for Novelty Alone. Evolutionary Computation 2011. [2] G. Vouros. Explainable Deep RL: State of the Art and Challenges. ACM Comput. Surv. 2022. [3] S. Milani et al. Explainable RL: A Survey and Comparative Review. ACM Comput. Surv. 2024. [4] L. Saulières et al. A Survey of Explainable RL: Targets, Methods and Needs. arXiv 2025. [5] N. Puri et al. Explain Your Move: Agent Actions via Specific and Relevant Feature Attribution. ICLR 2020. [6] Z. Juozapaitis et al. Explainable RL via Reward Decomposition. IJCAI/ECAI XAI Workshop 2019. [7] S. Tsirtsis et al. Counterfactual Explanations in Sequential Decision Making Under Uncertainty. NeurIPS 2021. [8] S. Sreedharan et al. Bridging the Gap: Post-Hoc Symbolic Explanations for Inscrutable Representations. ICLR 2022. [9] A. Verma et al. Programmatically Interpretable RL. ICML 2018. [10] O. Bastani et al. Verifiable RL via Policy Extraction. NeurIPS 2018. [11] S. Huang et al. Enabling Robots to Communicate their Objectives. Autonomous Robots 2019. [12] I. Bica et al. Learning “What-if” Explanations for Sequential Decision-Making. ICLR 2021. [13] G. Dao et al. Deep RL Monitor for Snapshot Recording. IEEE ICMLA 2018. [14] J. Wang et al. DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. IEEE TVCG 2019. [18] E. Leurent. An Environment for Autonomous Driving Decision-Making. GitHub 2018. github.com/eleurent/highway-env.